Unsupervised Anomaly Detection for a Smart Autonomous Robotic Assistant Surgeon (SARAS) Using a Deep Residual Autoencoder

Dinesh Jackson Samuel and Fabio Cuzzolin

***Abstract*—Anomaly detection in Minimally-Invasive Surgery (MIS) traditionally requires a human expert monitoring the procedure from a console, whereas automated anomaly detection systems in this area typically rely on classical supervised learning. Anomalous surgical events, however, are rare, making it difficult to capture data to train a model in a supervised fashion. In this work we propose an unsupervised approach to anomaly detection for robotic MIS based on deep residual autoencoders. The idea is to make the autoencoder learn the 'normal' distribution of the data and detect abnormal events deviating from this distribution by measuring a reconstruction error. The model is trained and validated upon both the publicly available Cholec80 dataset and a set of videos captured on procedures using artificial anatomies ('phantoms') as part of the Smart Autonomous Robotic Assistant Surgeon (SARAS) project. The system achieves recall and precision equal to 78.4%, 91.5%, respectively, on Cholec80 and of 95.6%, 88.1% on the SARAS phantom dataset. The system was developed and deployed as part of the SARAS platform for real-time anomaly detection with a processing time of 25 ms per frame.**

***Index Terms*—Surgical Robotics: laparoscopy, multi-robot systems, computer vision for medical robotics.**

I. Introduction

In RECENT years, Minimally-Invasive Surgery (MIS) has attracted a great deal of interest, as it only requires small incisions (5-30 mm) to provide the endoscope and other instruments access to the surgical cavity, rather than the vast ones (approximately 300 mm) demanded by traditional surgery. Endoscopic surgery results therefore in shorter recovery times compared to 'open' surgery. Although robotic MIS (R-MIS) technology is adaptive, precise and accurate, most R-MIS systems are not designed to replace the main surgeon conducting the procedure but to increase the safety and effectiveness of surgeries. One such kind of human-machine interactive robotic system, named 'da Vinci,' has been developed by Intuitive Surgical to perform precise and complex surgeries through small incisions [1]. Robotic-assisted surgical techniques have the potential to overcome human errors, by delivering high precision, reliability, and accuracy under human supervision. In fact surgical procedures can be very long, exhausting and cumbersome, leading to fatigue and hand trembling [2].

The typical surgical environment encompasses a patient table, a main surgeon, two assistant surgeons and two nurses. The assistant surgeon plays a key role both before and during the surgery. When using da Vinci, the main surgeon monitors and controls the robotic arm from the endoscopic console, while the assistant surgeon directs the da Vinci in handling the tools. These robots are not autonomous and can be considered as mere extensions of the main surgeon. Crucially, the assistant surgeon is active for only 30% of the time and remains idle during the rest of the surgery. As a result, using a da Vinci does not alleviate human surgeons' schedules, not does it lower the average cost of a surgical procedure. In addition, in a pandemic such as the one caused by Covid-19, severe shortages of assistant surgeons and additional safety measures gravely limit the number of surgeries to be carried out, at a cost of thousands of valuable lives.

A. Anomaly Detection in Endoscopic Surgery

During endoscopic minimally-invasive surgery, a surgical instrument known as *trocar* (basically a hollow tube) is inserted into a hole in the patient's body as a means of introduction for cameras and laparoscopic hand instruments. The main surgeons performs the surgery with the help of the visuals from the endoscopic camera; hence, a clear field of view is of paramount importance. Unfortunately, the latter often happens to be occluded in unexpected fashions, resulting in a limited usability of the endoscopic feed. Occlusions or any event that deviates from the normal workflow of the procedure are called *anomalies*. Different types of anomalies exist, including bleeding, the presence of smoke due to surgical electro-cauterisation, the blurring of camera lenses, or the camera being out of the trocar at some point during the surgery. Fig. 1 shows examples of anomalous scenes occurring during endoscopic surgery. In an autonomous surgical environment, in addition, anomalies may result from the camera being too close to the region of interest, tools completely occluding the view during the procedure or glares produced by organs or other tissue.

Traditional methods for detecting anomalies in endoscopic data combine traditional image-based feature extraction with

Manuscript received February 8, 2021; accepted July 1, 2021. Date of publication July 14, 2021; date of current version July 29, 2021. This work was supported by European Union's Horizon 2020 Research and Innovation Programme, under Grant 779813 (SARAS). This letter was recommended for publication by Associate Editor E. De Momi and P. Valdastri upon evaluation of the reviewers' comments. P. Valdastri. *(Corresponding author: Dinesh Jackson Samuel.)*

Dinesh Jackson Samuel is with the Postdoctoral Researcher and Associate Lecturer, Visual Artificial Intelligence Laboratory, Oxford Brookes University, Oxford OX33 1HX, U.K. (e-mail: rsamuel@brookes.ac.uk).

Fabio Cuzzolin is with the Oxford Brookes University, Oxford OX33 1HX, U.K. (e-mail: fabio.cuzzolin@brookes.ac.uk).

Digital Object Identifier 10.1109/LRA.2021.3097244

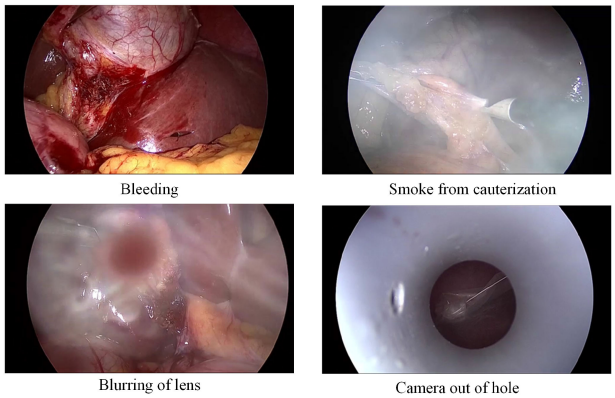

Fig. 1. Types of anomalies possibly occurring during cholecystectomy.

supervised learning [3]–[6] and mostly focus on bleeding. Such models are trained using both available anomalous and non-anomalous data. Liu and Yuan, for instance, employ colour-based feature extraction and a support vector machine (SVM) classifier to detect bleeding from images acquired through Wireless Capsule Endoscopy (WCE) [7]. Ghosh *et al.* have proposed a YIQ (Y-Luminance, IQ-Chrominance) color scheme for feature extraction from WCE videos. To differentiate bleeding from non-bleeding frames, statistical measures of the pixel values such mean, skewness, median, and minima are computed. Finally, a standard SVM is employed for classification. Similar approaches can be found in [8] and [9]. Research on smoke detection has been rather limited in endoscopy, whereas it has attracted more interest in non-medical applications such as forest fire and surveillance smoke detectors. Nevertheless, Leibetseder *et al.* have proposed a saturation analysis for extracting features from endoscopic video frames and used an 8-layer AlexNet for classification [10].

All such methods, however, go nowhere near as far as to meet the needs of real-time anomaly detection in R-MIS. The fundamental reason is that anomalous events of any kind, rather than the most anticipated ones such as bleeding and smoke, can happen in autonomous robotic surgery, making any approach based on supervised learning unsuitable. To compound the issue, to the best of our knowledge there currently is no accepted benchmark in anomaly detection in endoscopy, either supervised or unsupervised. Authors use different training/testing splits (e.g. a random selection of bleeding vs non-bleeding frames) and no dataset specifically designed for this task has been proposed yet.

B. Rationale for Unsupervised Deep Anomaly Detection

Real-time video anomaly detection in R-MIS is challenging due to the varied nature of the anomalies, the sparse occurrence of anomalous events, and the imbalance between the amount of data available under normal and abnormal conditions. Labelled data for anomalous events is typically unavailable and even hard to define in robotic-assisted surgery. A possible solution is provided by unsupervised learning, in particular in a deep learning formulation, which does not rely on any annotation. In particular, a class of deep neural networks known as *deep autoencoders* (DAE) has recently been proposed that is suitable for this task. An autoencoder consists of an encoder and a decoder. The encoder takes the input images/frames and reduces their dimensionality by mapping them to a latent space. The resulting 'bottleneck' features from this latent space are used by the decoder to minimise the error between original and reconstructed frames. As argued in [11], when applied to videos the learned autoencoder reconstructs regular motion with low error but incurs higher reconstruction error for irregular motions. The intermediate layers, whose size is smaller than both the input and the output layers, are designed for learning compact semantics as well as reducing noise. In fact, dimensionality reduction has been shown in the past to improve the result of other forms of unsupervised learning such as clustering [12]. Following this rationale, Jefferson *et al.* have proposed an anomaly detection system for videos which uses autoencoders and predictive convolutional LSTMs. Their model generates video frames from a sequence of input ones and predicts abnormal video frames using the reconstruction error principle [13].

Some efforts in the context of unsupervised video anomaly detection have been made in the wider computer vision field. E.g., Cho *et al.* have proposed an implicit autoencoder with a SlowFast network structure for anomaly detection in surveillance videos [14]. Recently, a robust unsupervised anomaly detection approach has been proposed by Wang *et al.* which employs a ConvGRU-based prediction network for capturing the spatiotemporal dependencies characterising normal data to predict anomalous frames [15].

C. Contributions

In this letter we propose a deep autoencoder-based approach to real-time video anomaly detection, which learns spatial information from the input video frames and strives to reconstruct the input video frames with low error. The approach is validated using two datasets: the publicly available Cholec80 dataset, adapted for anomaly detection, and a SARAS dataset of videos capturing radical prostatectomy (RARP) procedures conducted on artificial anatomies (*phantoms*), achieving extremely promising results. In summary:

- We propose the first deep unsupervised system for video anomaly detection in endoscopic surgery.
- We contribute the first benchmarks for unsupervised surgical anomaly video detection, based on the existing Cholec80 dataset and our own SARAS dataset.
- Our results show that our approach is able to detect all typical surgical video anomalies with high accuracy.
- The system has been implemented and deployed in SARAS for video anomaly detection in real-time.

We plan to release our annotation upon acceptance to share the new benchmarks with the community.

II. Deep Residual Autoencoders

Unsupervised learning approaches have recently gained momentum in computer vision, thanks for their not relying on expensive and time-consuming labeled datasets. Here we propose

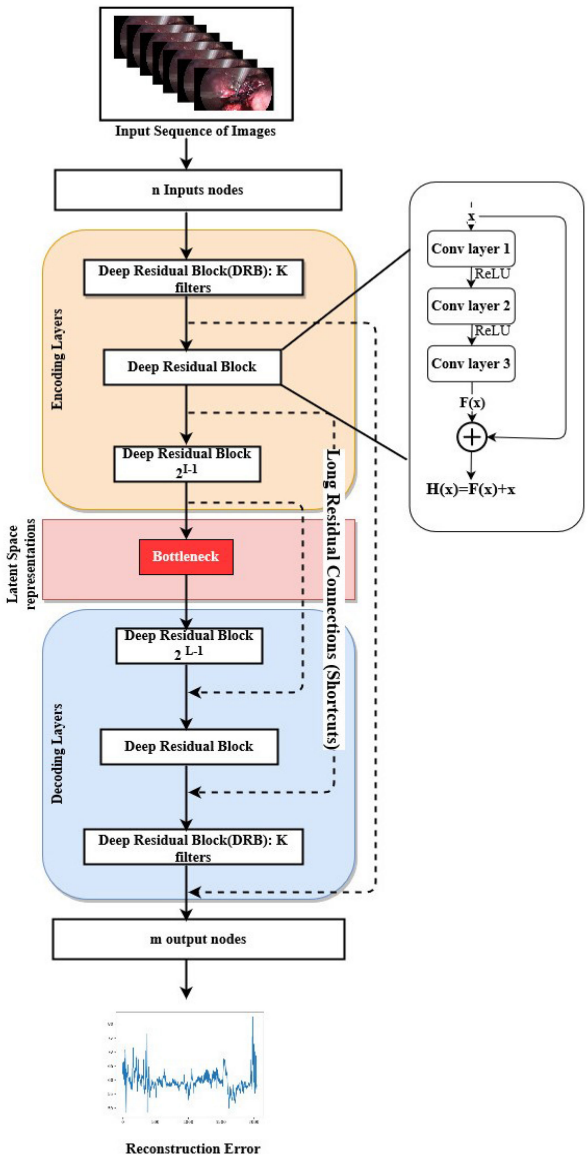

Fig. 2. Architecture of a deep residual autoencoder.

in particular to tackle anomaly detection in an unsupervised approach based on deep residual autoencoders.

The latter are a form of generative deep neural network, inspired by the discovery by neuroscience of shortcut connections in the brains of various animals, in turn emulated by residual convolutional neural networks [16], [17]. The term *residual learning* relates to variables that consist of residual vectors between two segments of a long sequence. Residual vectors have been shown to be effective in learning shallow feature representations for image recognition tasks [15], [16]. Based on these facts, *residual convolutional neural networks* have been proposed for improving the accuracy of deep learning networks [18], [19].

The architecture consists of a continuous, stacked sequence of deep residual blocks (DRB), connected in a sequence with shortcut connections in a residual CNN. Each deep block consists of three convolution layers (see Fig. 2). The three sequentially-connected convolution layers collectively map an input x to an output $F(x)$. The latter is then added to the input x of the DRB via a shortcut connection. Overall, the output of the deep residual block can be expressed as $H(x) = F(x) + x$. The condition $F(x) = 0$ indicates the disappearance of the network gradient weights, in which case $H(x) = x$ tends to an identity mapping that decreases the network's depth while guaranteeing classification accuracy [20]. Biological findings in the brain show the pivotal role of similar hidden shortcut connections for synchronised motor movement, recovery from injuries, and reward learning. This mechanism has thus been injected into deep learning, namely residual CNNs and U-Net [21], [22].

III. Methodology

The proposed methodology for anomaly detection is thus based on an unsupervised learning approach using a deep residual autoencoder. The shallow layers of the encoder are connected to the decoder's deep layers using shortcuts to encourage the formation of identity mappings (see Fig. 2 again). The autoencoder learns from a dataset of 'normal' videos and uses the learned parameters to identify abnormal behavior by thresholding the reconstruction error. When compared with the concatenated shortcuts in U-Net [22], these residual connections have the property of minimising the number of model parameters and of enhancing learning by propagating gradients between layers more efficiently.

A. Architecture

A typical deep residual autoencoder has n inputs and m target outputs. The encoder part consists of a sequence of Deep Residual Blocks which aim to compress the input image to a latent representation. This is followed in the decoder by a sequence of the same number of DRBs designed to reconstruct an input image by taking it back to its original dimension (Fig. 2). The architecture thus has a symmetrical structure in which layers laying on opposite sides of the bottleneck mirror each other.

Each deep residual block comprises three convolutional layers, followed by a rectified linear unit (ReLU) activation function and batch normalisation (BN) for re-scaling to improve training. Average pooling is applied after the convolutional layers to extract rich, smooth spatial features while preserving localisation information. Because of this, the output of each residual block differs in terms of the size of the resulting feature map, making it difficult for the gradients to propagate through consecutive blocks at training time. To address this problem two types of residual connections, known as 'short' and 'long' connections are used in the model. A short residual connection is used locally within the convolutional layers of each DRB, where feature maps have identical dimensions across layers. Long residual connections are also established between pairs of DRBs in corresponding encoder and decoder layers. These cut across the bottleneck to tackle the issue with vanishing gradients at a global, network level.

More in detail, let l denote an encoding layer and L the corresponding decoding layer of the network. The input and output of the encoding layer l are denoted by X_l and Y_l, respectively, and by X_L and Y_L for the corresponding decoding layer L. The

residual connection between corresponding layers mitigates the loss of information when backpropagating losses during training [23]. The relationship among the relevant quantities is illustrated in Eq. (1):

$$Y_L = X_l + f_L(X_L, W_L) = X_l + f_L(g_L(f_l(X_l, W_l)), W_L), \tag{1}$$

where the activation functions for the encoding layer l and the corresponding decoding layer L are denoted by $f_l(X_l, W_l)$ and $f_L(X_L, W_L)$, respectively. Consequently, $X_L = g_L(f_l(X_l, W_l))$ represents the recursive mapping between the shallow encoder layer input X_l and the deep decoder layer input X_L. In Fig. 2 K denotes the size of the convolution kernels. Deep autoencoders can efficiently learn an encoding from input data through dimensionality reduction via bottleneck features [24].

B. Reconstruction Error-Based Detection

As explained above, in our approach anomaly detection is reconstruction error-based. The hypothesis is that, after the learning process, the autoencoder can reconstruct input frames never seen before, under the assumption that the latter resemble 'regular' frames observed during training. Conversely, an autoencoder would struggle to reconstruct anomalous frames not matching the learnt feature maps. Therefore, abnormal frames would have a high reconstruction cost when compared to normal frames. Our error measure is based on the relative difference between normal and abnormal structures, rather than its absolute value. To evaluate the performance of the system, we use a robust regressive loss function known as *root mean squared error* (RMSE). The RMSE loss function heavily penalises reconstructed values which stray far from 'normal' values, and is defined as:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{M}\sum_{j=1}^{N}\left(R\left[i,j\right] - F\left[i,j\right]\right)^2}{M \times N}}, \tag{2}$$

where $R[i, j]$ and $F[i, j]$ are the pixel values of the reference image coordinates and the reconstructed image, respectively.

IV. Experimental Results

Our residual autoencoder architecture was trained on a Quadro RTX 6000 8-GPU server with 24 GB VRAM per card. Performance was evaluated using both anomalous and non-anomalous (normal) frames generated by the SARAS demonstration platform and an existing dataset in the surgery domain, Cholec80, adapted for anomaly detection. In our tests, each frame in the input video was passed to the autoencoder in order to measure the reconstruction error.

A. Datasets

The Cholec80 dataset contains 80 videos of cholecystectomy procedures, performed by 13 different surgeons [25]. The videos are captured at 25 frames per second (fps) with a resolution of 854 $\times$ 480 pixels, and contain both anomalous and non-anomalous frames. As it was designed for phase recognition and tool detection rather than anomaly detection, in Cholec80 each frame is labelled with the phase of the procedure and the presence of tools (at 1 fps). The model was also tested upon a dataset acquired through the SARAS SOLO-SURGERY system [26]. Prostatectomy was performed on a 3D-printed 'phantom' prostate by both a da Vinci and a pair of SARAS robotic arms. Our SARAS dataset contains 18 video clips captured at 25 fps and 720 $\times$ 480 resolution, portraying only non-anomalous data (as anomalies are difficult to simulate in a phantom environment). SARAS videos also come with extra annotation in the form of 23 relevant classes of surgeon actions, as in the real-world SARAS-ESAD surgical dataset used for a recent MIDL 2020 challenge on surgeon action detection [27]. Note that, although the SARAS dataset does not contains any *surgical* anomalies, it is affected by a number of *technical* ones such as blocked views or loss of focus, which can actually mimic anomalies one would expect in a real-world setting.

B. Training and Testing

Selection. In our tests the data for training the model is carefully chosen to avoid the presence of any anomalous frames. This does amount to a sort of 'implicit' supervision signal, making the distinction between supervised and unsupervised approaches less distinct. In each dataset we selected a subset of frames for training and a separate, disjoint set of frames for testing. For Cholec80 we identified 65 'normal' video clips for training. For testing, 3 video clips containing anomalies were selected using domain-specific knowledge from surgeons, as this is crucial for this kind of work. In total 5584 frames were selected for testing and manually labelled. The model for the SARAS phantom dataset was trained upon a suitable selection of frames from the 18 video clips and tested on a total of 4799 frames. Again, domain knowledge was used in the selection process.

Protocol. For training, video frames were resized to 128 $\times$ 128 pixels to be passed as input to the deep residual encoder-decoder pipeline. While a larger batch size and learning rate may affect the model's generalisation power due to sharp swings in the training function, a smaller batch size helps the model converge faster to a better global optima. Hence, in our experiments we set an optimal batch size of 512 frames and a learning rate of 0.001. An Adam optimizer was used. As a loss function we adopted L_2 Mean Square Error (MSE), which measures the average square error between input and reconstructed images. The model was validated using a small fold containing 583 normal and 442 anomalous frames in order to find the best number of epochs for model training. The accuracy on the validation fold was computed for different values of the number of epochs to select the best such value, resulting in 40 epochs for Cholec80 with an anomaly detection accuracy of 85.2% and in 30 epochs on the SARAS dataset with an accuracy of 94.6%. In our tests we found that it was not necessary to introduce all types of anomalies in the validation fold, as even a small fraction of anomalous frames is enough to make the model learn the data distribution and find the best hyperparameters.

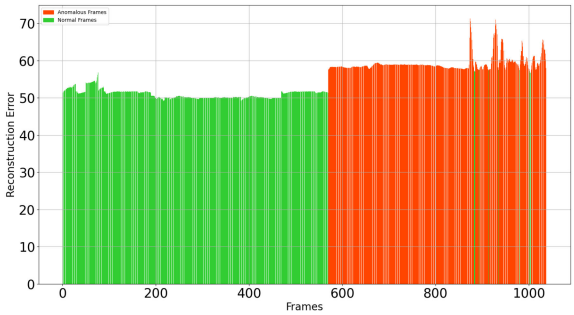

Fig. 3. Reconstruction error distribution over SARAS's test fold.

C. Anomaly Detection

Once the optimal epoch number is identified, the model is tested on the selected test fold, which contains both anomalous and non-anomalous frames.

Error distribution. Fig. 3 shows the distribution of the reconstruction error over the test fold for the SARAS dataset, as produced by the trained autoencoder. The graph shows a clear separation between the error associated with the two types of frames, supporting the validity of our reconstruction-based approach.

Choice of threshold(s). To quantitatively discriminate anomalies and prompt the system to take precautionary actions, such as suction in case of bleeding or smoke or calling for manual intervention for other types of anomalies, we employ a threshold approach. The optimal threshold is found using the $n-th$ percentile of the error distribution from normal frames. For instance, $n = 95$ means that the threshold is such that 95% of the normal samples have a reconstruction error lower than the threshold. For SARAS this yields 57.4, so frames whose reconstruction error is above 57.4 are considered as anomalous. To find the optimal value of n we ran an empirical analysis by plotting the accuracy as a function of n and selecting the optimal n^*.

For the Cholec80 dataset we selected both a lower threshold θ_i and an upper threshold θ_j, so that an anomaly would be flagged whenever the reconstruction error for the test frame was below θ_i or above θ_j. The reason for using a lower and an upper threshold is the effect of smoke, which tends to lower the value of the reconstruction error due to saturation of the pixel intensities in the frame. The upper threshold was determined using the $n-th$ percentile approach, while the lower one was set to the lowest reconstruction error for a normal frame.

Fig. 4 shows the interactive visualisation dashboard we deployed for live data experimentation, with both the reconstruction error graph generated by the trained autoencoder and the reconstructed images for some sample anomalous frames. The X-axis represents the frame number, while the Y-axis represents the reconstruction error value.

Performance measures. After having an expert manually label a significant fraction of the video frames in the two datasets as either normal or anomalous, as already explained, our anomaly detection approach was evaluated by means of the usual precision and recall measures:

$$Recall = \frac{TP}{TP+FN}, \quad Precision = \frac{TP}{TP+FP}, \qquad (3)$$

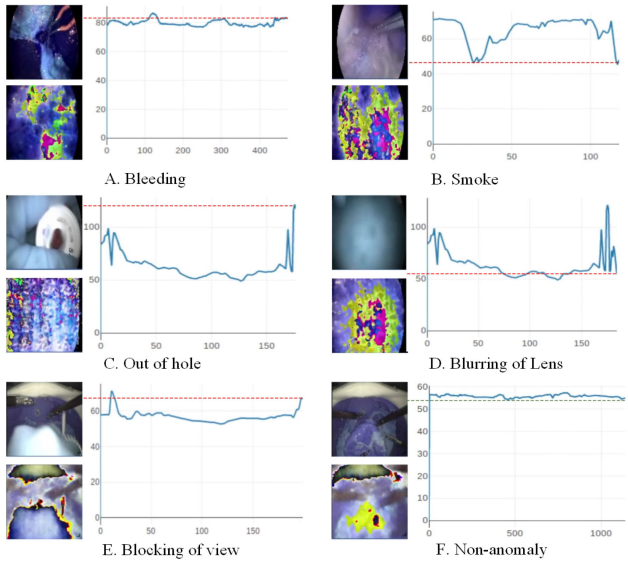

Fig. 4. A, B, C and D plot the reconstruction error over time and a pair of reconstructed/actual anomalous frames for four videos in the Cholec80 dataset. The X-axis represents the frames and Y-axis represents the reconstruction error. Anomalies with different causes are considered. E and F do the same for two videos of the SARAS dataset, one containing an anomaly (blocked view, corresponding to spikes in the reconstruction error) and one not containing any. In each block the frame portrayed on the left is the last in the sequence plotted in the graph.

TABLE I
RECALL, PRECISION, F1-SCORE OVER THE CHOLEC80 TEST FOLD

Anomalies	No. frames	Recall	Precision	F1
Ganomaly (overall)	5,584	57.5	64.2	60.6
DAE without RC (overall)	5,584	72.6	83.5	77.6
Our model (overall)	5,584	**78.4**	**91.5**	**84.4**
Our model (bleeding)	2,925	76.6	93.7	84.2
Our model (smoke)	411	100	89.2	94.2
Our model (camera out of hole)	754	76.6	89.2	71.3

TABLE II
RECALL, PRECISION, F1-SCORE OVER THE SARAS TEST FOLD

Model	No. frames	Recall	Precision	F1
Ganomaly	4,799	61.3	69.4	65.1
DAE without RC	4,799	86.2	79.6	82.7
Our model	4,799	**95.6**	**88.1**	**91.6**

where TP is the number of true positives (anomalous frames), FP the number of false positives and FN that of false negatives (normal frames). Although anomaly detection is completely unsupervised, this labelling was conducted to allow a quantitative assessment of the system's performance.

D. Detection Accuracy

In Tables I and II the performance of our model is compared with that of two competitors: the *Ganomaly* (generative adversarial network for anomaly detection) approach in [28] and a deep autoencoder similar to ours but without residual connections (*DAE without RC*). Standard evaluation

metrics such as recall, precision and F1-score are used for performance analysis on the Cholec80 and SARAS dataset, respectively.

As it can be appreciated, the Ganomaly [28] approach of estimating the reconstruction error directly in the latent space using a GAN architecture turns out to be inferior to our approach of measuring the reconstruction error in the RGB space. Furthermore, autoencoders without residual connection are shown not to be good at reconstructing frames, because of the aforementioned issue with gradient loss in the encoding/decoding cascade. The absolute performance of our model is quite satisfactory, although it flags significant challenges. On Cholec80, our model detects anomalies with a specificity of 57.1%, a recall of 78.4% and a precision of 91.5% using a threshold range of $[\theta_i = 48, \theta_j = 65]$. On SARAS, our model achieves a specificity of 98.5%, a recall of 95.63% and a precision of 88.10%. Here anomalies are detected using a upper threshold value $\theta_i >= 57.4$.

Speed-wise, the processing time on a single NVIDIA GTX 1070 GPU with 8 GB VRAM is 25 ms per frame, for clips with frame resolution 854×480.

V. Discussion and Conclusions

Overall, the proposed architecture has been shown capable of detecting anomalies in basically real-time in an unsupervised fashion and with satisfactory accuracy. The model uses deep autoencoders with residual connections to propagate gradient values throughout the network more reliably. The complete system was validated on the real-data Cholec80 surgical dataset (suitably augmented) and the SARAS phantom dataset, achieving promising results for unsupervised anomaly detection in endoscopic videos.

We compared our results with that of a GAN based approach in which error is measured in the latent space, and with that of autoencoders without residual connections, showing the superiority of our method. Nevertheless, the absolute performance numbers attest that much progress is needed in unsupervised methods for anomaly detection to allow their real-world deployment in autonomous surgical environments. A number of research avenues remain open. One could think of adding a small quantity of anomalies to the validation fold (since annotation is available) to enforce separation. However, in most cases there are no anomalies to do supervised learning upon (e.g. in the SARAS phantom dataset surgical anomalies are not present). In addition, if we teach the system what an anomaly is during training it will likely not be able to detect other (possibly unforeseeable) forms of anomaly not in the training set.

Another observation is that the reconstruction error may vary even in absence of anomalies depending on the complexity of the image. Furthermore, some anomalies are spatially localised (e.g. bleeding), suggesting that the overall reconstruction error should be replaced by a pixel-wise distribution of the error. We will explore autoencoder architectures which exploit this principle in the near future.

Finally, the approach can be extended to detect anomalies using 3D video feature extraction in a GAN architecture, combining the strength of the two approaches.

References

[1] G. S. Guthart and J. K. Salisbury, "The Intuitive/sup TM/telesurgery system: Overview and application," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2000, pp. 618–621.

[2] S. L. Lee *et al.*, "From medical images to minimally invasive intervention: Computer assistance for robotic surgery," *Comput. Med. Imag. Graph.*, vol. 34, no. 1, pp. 33–45, 2010.

[3] Y. Fu, W. Zhang, M. Mandal, and M. Q. Meng, " Computer-aided bleeding detection in WCE video," *IEEE J. Biomed. Health Informat.*, vol. 18, no. 2, pp. 636–642, Mar. 2014.

[4] B. Li and M. Q. Meng, "Computer-aided detection of bleeding regions for capsule endoscopy images," *IEEE Trans. Biomed. Eng.*, vol. 56, no. 4, pp. 1032–1039, Apr. 2009.

[5] X. Xing, X. Jia, and M. Q.-H. Meng, "Bleeding detection in wireless capsule endoscopy image video using Superpixel-Color histogram and a subspace KNN classifier," in *Proc. 40th Annu. Int. Conf. IEEE Eng Med. Biol. Soc.*, 2018, pp. 1–4.

[6] Y. S. Jung, Y. H. Kim, D. H. Lee, and J. H. Kim, "Active blood detection in a high resolution capsule endoscopy using color spectrum transformation," *Proc. Int Conf BioMed. Eng. Informat.*, 2008, pp. 859–862.

[7] J. Liu and Y. Xiaohui, "Obscure bleeding detection in endoscopy images using support vector machines," *Opti. Eng.*, vol. 10, pp. 289–299, 2009.

[8] S. Sainju *et al.*, "Automated bleeding detection in capsule endoscopy videos using statistical features and region growing," *J. Med. Syst.*, vol. 38, no. 4, pp. 25–36, 2014.

[9] T. Okamoto *et al.*, "Real-time identification of blood regions for hemostasis support in Laparoscopic surgery," *Signal, Image Video Process.*, vol. 13, pp. 405–412, 2019.

[10] A. Leibetseder *et al.*, "Real-time image-based smoke detection in endoscopic videos," in *Proc. Workshops ACM Multimedia*, 2017, pp. 296–304.

[11] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis, "Learning temporal regularity in video sequences," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* 2016, pp. 733–742.

[12] F. Cuzzolin, D. Mateusy, D. Knossow, E. Boyer, and R. Horaud, " Coherent Laplacian 3-D protrusion segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2008, pp. 1–8.

[13] J. Medel and A. Savakis, "Anomaly detection in video using predictive convolutional long short-term memory networks," 2016, *arXiv:cs/1612.00390*.

[14] M. Cho *et al.*, " Unsupervised video anomaly detection via flow-based generative modeling on appearance and motion latent features," 2020, *arXiv:cs/2010.07524*.

[15] X. Wang *et al.*, "Robust unsupervised video anomaly detection by multipath frame prediction," 2020, *arXiv:cs/2011.02763*.

[16] A. Saunders *et al.*, "A. direct GABAergic output from the basal ganglia to frontal cortex," *Nature*, vol. 521, no. 7550, pp. 85–89, 2015.

[17] M. Zelikowsky *et al.*, "Prefrontal Microcircuit underlies contextual learning after hippocampal loss," *Proc. Nat. Acad. Sci.*, vol. 110, no. 24, pp. 9938–9943, 2013.

[18] H. Jégou, F. Perronnin, M. Douze, J. Sánchez, P. Pérez, and C. Schmid, "Aggregating local image descriptors into compact codes," *IEEE Trans. Pattern Anal. Mach. Intell.* vol. 34, no. 9, pp. 1704–1716, Sep. 2012.

[19] R. Szeliski, "Locally adapted hierarchical basis preconditioning," in *Proc. ACM SIGGRAPH Papers* 2006, pp. 1135–1143.

[20] A. Veit, M. Wilber, and S. Belongie, "Residual networks behave like ensembles of relatively shallow networks," 2016, *arXiv:cs/1605.06431*.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* 2016, pp. 770–778.

[22] O. Ronneberger, P. Fischer, and T. Brox,, " U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Med. Image Comput. Comput. Assist. Interv.*, 2015, pp. 234–241.

[23] L. Li, "Deep residual autoencoder with Multiscaling for semantic segmentation of land-use images," *Remote Sensing*, vol. 11, no. 8, 2019, Art. no. 2142.

[24] C. Y. Liou *et al.*, "Autoencoder for polysemous word," *Lecture Notes in Computer Science.* Berlin, Heidelberg: Springer, vol. 7751, 2013.

[25] A. P. Twinanda, S. Shehata, D. Mutter, J. Marescaux, M. de Mathelin, and N. Padoy, "EndoNet: A deep architecture for recognition tasks on laparoscopic videos," *IEEE Trans. Med. Imag.*, vol. 36, no. 1, pp. 86–97, Jan. 2017.

[26] A. Leporini *et al.*, "Technical and functional validation of a teleoperated multirobots platform for minimally invasive surgery," *IEEE Trans. Med. Robot. Bionics*, vol. 2, no. 2, pp. 148–156, May 2020.

[27] V. S. Bawa *et al.*, "ESAD: Endoscopic surgeon action detection dataset," 2020, *arXiv:cs/2006.07164*.

[28] S. Akcay *et al.*, "GANomaly: Semi-supervised anomaly detection via adversarial training," in *Proc. Asian Conf. Comput. Vis.* 2018, pp. 622–637.